Improving the Presumptive Disability Decision-Making Process for Veterans

Committee on Evaluation of the Presumptive Disability Decision-Making Process for Veterans, Jonathan M. Samet and Catherine C. Bodurow, Editors

ISBN: 0-309-10731-8, 440 pages, 6 x 9, (2008)

This free PDF was downloaded from:
http://www.nap.edu/catalog/11908.html

THE NATIONAL ACADEMIES
Advisers to the Nation on Science, Engineering, and Medicine

8

Synthesizing the Evidence for Causation

In Chapter 7 we discussed the challenges of inferring a causal relationship between an exposure and a health outcome, and the range of evidence typically considered in making such inferences. In this chapter we discuss the problem of combining potentially diverse types of evidence in making a single, overall judgment about whether an exposure causes a health outcome. We begin by discussing the problem of integrating the evidence from multiple epidemiologic studies. We then describe a framework for combining epidemiologic and other evidence into a single *quantitative* judgment about the strength of causation. Next we discuss *qualitative* frameworks that have been used by expert committees for categorizing the overall strength of evidence for or against a causal claim. And lastly we propose a qualitative framework for causal reference to be used in the presumptive disability decision-making process.

We include this material because the scientific group that we propose in our new approach will review evidence on the health of veterans and will have the task of integrating an accumulating stream of results and interpreting new findings in the context set by previous findings and prior reviews. The successive Institute of Medicine (IOM) Agent Orange committees are illustrative. While the methods are presented in a theoretical fashion, they would be key components of the approach recommended by this Committee and, in reality, they are already inherent to the approaches used by the IOM committees.

META-ANALYSIS: COMBINING EVIDENCE FROM MULTIPLE STUDIES

Scientific evidence relevant to causal relationships between exposure and disease comes from different types of investigation, including randomized clinical trials (RCTs) on humans, epidemiologic studies, animal experiments, and cell studies, and also from fundamental biological knowledge. We use the term *human studies* to refer to RCTs or observational studies involving people. Although an evidence-based approach must combine all forms of scientific evidence, in this section we limit our discussion to the problem of synthesizing the information from multiple human studies.

The idea of pooling information from multiple studies has a long tradition in statistics that goes back at least to Karl Pearson in 1904. A meta-analysis involves gathering all studies with evidence related to a particular question, and statistically combining the results of these studies. In many contexts, health researchers have mathematically combined the results from multiple, yet comparable RCTs to derive a summary estimate of the effect of some substance on health; the estimate appropriately combines the results of all the individual studies. Such summaries are often carried out, for example, to determine if there is a benefit of a drug or perhaps an excess occurrence of an unwanted side effect. One approach for combining evidence, random effects meta-analysis, allows for heterogeneity between studies; with this technique, a meta-analysis is not strictly limited to studies involving similar populations.

In observational studies, there may be more variability in findings from study to study because study variables are not under the investigator's control. The populations studied may vary considerably in their characteristics, and the variables measured as covariates for statistical adjustment may also differ. Nevertheless, meta-analysis is applied to observational study results as well as to RCT data. Meta-regression (Greenland and O'Rourke, 2001) allows pooling of data across observational studies with some unexplained heterogeneity, and recent work by E. Kaizar (2005) improves on meta-regression for situations with data available from both RCTs and observational studies.

Although the development of meta-analytic methods has generated extensive methodological discussion (see, for example, Berlin and Antman, 1994; Berlin and Chalmers, 1988; Dickersin and Berlin, 1992; Greenland, 1994a,b; Stram, 1996; Stroup et al., 2000), it is a technique that can be quite useful when there are multiple studies on the same question. For example, for each of a number of different cancers, the 2006 IOM Committee on Asbestos and Selected Cancers (IOM, 2006a) did a quantitative meta-analysis on studies that combined the effect of asbestos exposure on risk based on multiple studies for each of a set of cancers. The report pre-

sented the results of individual studies as well as an overall estimate that came from the combination of the estimates from the individual studies.

THE BAYESIAN APPROACH

No matter how sophisticated the meta-analytic technique, it is still limited to combining statistical evidence from different studies into a single statistical estimate of the effect size. As we discussed in Chapter 7, the scientific evidence germane to causal claims also includes mechanistic knowledge, findings of animal or cell and molecular studies, and other knowledge relevant to biological plausibility. A technique for combining all the available evidence into a single judgment needs to accommodate these other types of information. One approach for combining all the evidence available into a single quantitative judgment uses a *Bayesian* approach.

Bayesian methodology conceives of probability as degrees of belief. Any proposition can be given a degree of belief. For example, one person might have a personal degree of belief of 0.30 (30 percent) in the proposition that garlic prevents colds, while another might give the proposition 0.80. The Bayesian approach provides a rule for updating the existing degree of belief in response to additional evidence (Bayes' rule, see below). Provided that no experts are inflexible in their belief,[1] a group of experts who update their beliefs by Bayes' rule will almost certainly converge to the same degree of belief after considering enough of the same relevant evidence.[2] Thus, a group of scientists considering the overall evidence for a causal claim such as "formaldehyde causes leukemia" might begin with a *prior* degree of belief about the claim and then update their belief in the light of accumulating evidence, regardless of the type of evidence. In more detail, if we consider each separate causal model relating formaldehyde (exposure) and leukemia as a separate hypothesis, H_i, then we can begin by using background knowledge or the results of previous studies, to assign a *prior probability* to each such hypothesis.

The Bayesian approach then seeks to compute the *posterior probability* $\Pr(H_i \mid D)$, namely the probability of (i.e., degree of belief in) hypothesis H_i given the new observed data D. The computation of these posterior probabilities is given by the famous rule first described by the Rev. Thomas Bayes, known as Bayes' theorem or rule, which in simple form is

$$P\left(H_i \middle| D\right) = \frac{P\left(H_i\right) P\left(D \middle| H_i\right)}{P\left(D\right)}.$$

---

[1] That is, assigns degree of belief 0 to what turns out to be the "true" proposition.
[2] And a raft of other assumptions—see Howson and Urbach, 1989.

Loosely translated, this formula reflects the prior probability for the hypothesis $P(H_i)$, the *likelihood*, meaning the probability of the data D given the hypothesis H_i is true, and $P(D)$, the probability of the data before any hypothesis.

The Bayesian framework provides a useful perspective on how weak additional evidence may have little impact on a strong prior probability or strong additional evidence may have substantial impact on a weak prior probability. For example, assume that we have very strong prior radiobiological and epidemiologic evidence that ionizing radiation can cause cancer, but a study of veterans who participated in nuclear weapons testing maneuvers fails to provide strong evidence of an association. In this example, assume that the data from the study do not lead to rejection of the null hypothesis of no association in a study of veterans exposed to radiation. This null finding might arise because the available sample size or length of follow-up was too small to yield an adequate number of cancers in the study group or because of inherent biases in the study design, such as inaccuracy of the available radiation dose information. If the veterans' data are highly uncertain, then a null result might change the prior estimate of population radiation risk (e.g., the synthesis of the world literature by expert committees such as the National Research Council's [NRC] Biological Effects of Ionizing Radiation [BEIR] committee) downward for this group, but by only a very modest amount. The posterior probability would still strongly favor a presumption (Figure 8-1a).

On the other hand, assume that there was very little prior knowledge about the effect of dioxins on cancer risk and a large, well-designed study of Vietnam veterans yielded a large and highly significant positive relative risk for non-Hodgkin's lymphoma (NHL) (Figure 8-1b). In this case the absence of prior knowledge would have little influence on the degree of judgment given to the estimates from the veterans' study.

A further example is provided by the National Research Council's Biological Effects of Ionizing Radiation (BEIR) IV Committee (NAS, 1988), which addressed the cancer risk of plutonium in humans. The available human data were very limited: no bone cancers were observed among 18 patients injected with plutonium for experimental purposes or among a small number of others occupationally exposed in the Manhattan Project. The BEIR IV committee adopted an empirical Bayes approach that assumed that the ratio of carcinogenic potencies of plutonium to various other radionuclides would be roughly constant across species. As there are substantially more human data about the carcinogenicity of various isotopes of radium and extensive animal data about plutonium, radium, and other radionuclides, it was possible to estimate the risk of plutonium in humans from a combined analysis that incorporated animal and human data. The committee carried out an uncertainty analysis that incorporated the variability in the ratios
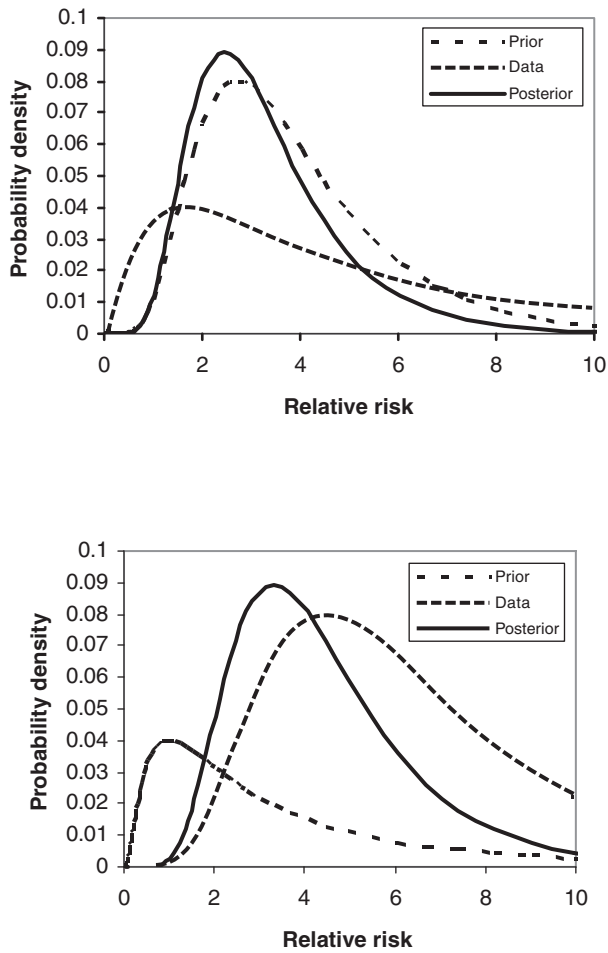
FIGURE 8-1 Hypothetical illustrations.
NOTE: Hypothetical illustrations of the combination of prior knowledge or judgment with study data to yield posterior estimates of a causal parameter, here the relative risk: (a) strong prior probability for a nonnull effect combined with weak data showing little or no effect in the study sample, and (b) weak prior probability combined with strong data showing a major effect.

of the relative carcinogenicities of the radionuclides across species. In this analysis, the limited human data for plutonium shifted the posterior probability distribution of the carcinogenicity estimates only slightly downward relative to the prediction resulting from the animal data.

The Bayesian approach can also focus scientific attention as new evidence becomes available. For example, as discussed in Chapter 7, there are many ways to generate an observed association between an exposure and a health outcome. The association might be the result of the exposure causing the health outcome, confounding, other forms of bias or of chance. In carrying out an observational study, epidemiologists try to measure and adjust for confounders and eliminate bias with good study design and appropriate data analysis. Scientific opinion about how much of the adjusted association might still be from unmeasured confounding or bias is useful in judging the degree of confidence about the estimate of an association.

To illustrate, consider the diagram in Figure 8-2, which graphically depicts the epidemiologists' concerns. Assuming a simple linear model, for illustration, the parameter β represents the *causal association*, that is, the amount of observed association between exposure and health due to the causal influence of exposure on health. The parameter α represents the amount of association from confounding that we *can* statistically adjust for,
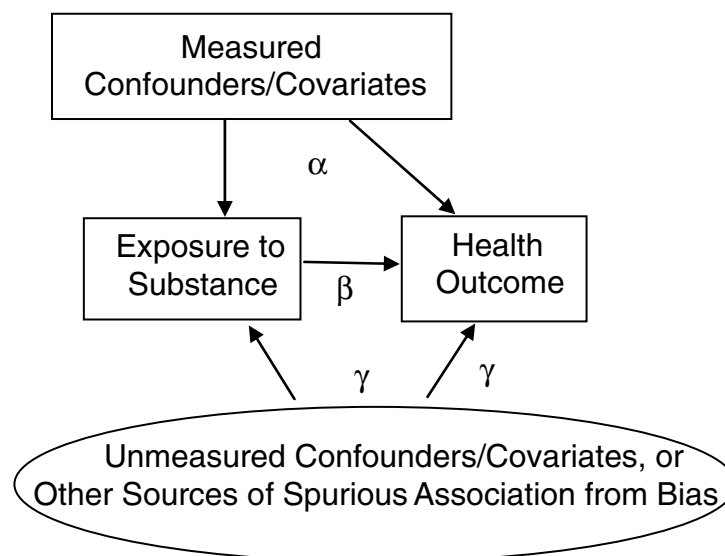

FIGURE 8-2 Focusing on unmeasured confounders/covariates, or other sources of spurious association from bias.

given the right model, and the parameter γ represents the amount of spurious association that we *cannot* statistically adjust for. Because the estimate of the causal association β will be biased in proportion to the size of γ, scientific attention should be focused on γ. In a Bayesian approach, we can encode scientific opinion about the size of γ into a prior probability, and then compute a posterior probability over β that appropriately takes into account uncertainty over γ.

In principle, then, the Bayesian approach provides an entirely quantitative framework for combining theoretical beliefs and evidence from previous studies along with the data at hand to update estimates of model parameters or of the probability that a particular hypothesis is true. In principle, each researcher undertaking a new study would apply the procedure to interpret the new evidence from the study in the context of already existing evidence, arriving at a posterior probability that is a methodologically appropriate combination of prior beliefs and accumulated evidence.

In practice, however, the Bayesian approach is *far* from a panacea for the complex task of combining diverse types of evidence. In typical scientific contexts, it may be difficult to move from fairly inchoate and diverse sorts of background knowledge to a communal "prior." By their nature, these prior odds are a matter of judgment, about which consensus amongst scientists can be difficult to obtain.

Further, although "updating" a posterior probability from certain kinds of evidence is reasonably straightforward, updating from other kinds of evidence is not. For example, consider estimating the relative risk of leukemia as a function of benzene exposure. After specifying a prior probability over this function, updating from a new sample of 400 veterans who were exposed to twice the usual level of benzene and who have a relative risk of 1.8 for leukemia is fairly straightforward mathematically. How are we to update based on experimental evidence showing that rats exposed to 50 times the background level of benzene develop leukemia at 3 times the rate of those exposed to background levels? What is the likelihood of this evidence, assuming any particular hypothesis about humans? Here we move beyond mathematics and statistics and into opinion about the comparability of leukemia and its causation in rats and humans.

QUALITATIVE FRAMEWORKS USED BY EXPERT COMMITTEES

Faced with diverse mechanistic and biological evidence that cannot be incorporated into a single statistical meta-analysis, and with opinions and judgments too varied and vague to employ a formal Bayesian approach, expert committees have resorted to qualitative categorizations of the strength of evidence for causation. There is a lengthy history of doing so, dating back to the 1950s as evidence began to develop on disease causa-

tion by radiation and tobacco smoking (Bayne-Jones et al., 1964). Judgments as to the level of evidence for causation can have substantial impact and often have regulatory implications. For example, the International Agency for Research on Cancer (IARC) (2006b), the U.S. Environmental Protection Agency (EPA) (2005), and the National Toxicology Program (NTP) (NTP, 2005) have developed systems for classifying the level of evidence in support of a causal relationship between chemicals and cancer. Similar classification systems have been developed for causal relationships for other health-specific effects such as reproductive outcomes (see e.g., NTP CERHR, 2003, 2005; Shelby, 2005), or between agents and health outcomes in general (IOM/NRC, 2005), or between smoking and disease (DHHS/CDC, 2004).

Each of these classification systems relies on evidence from a variety of research sources: epidemiologic, toxicological, and biological. The approach of combining diverse sources into an overall judgment on the strength of evidence for general causation, at least in a public health context, can be traced back to the 1964 *Report of the Advisory Committee to the Surgeon General on Smoking and Health* (Bayne-Jones et al., 1964), as well as other early summary reports on smoking and health. In the introductory chapters to the 1964 report, the committee described the different sorts of evidence to be considered. They specifically listed animal experiments, clinical and autopsy studies, and population studies, but were expansive in the evidence considered. They described the importance of expert evaluations of the quality of published reports, wrote two pages on their working definition of causation, and codified a subset of Sir Bradford Hill's criteria for establishing causation in epidemiology and public health contexts: consistency of the association, strength of the association, specificity of the association, temporality of the association, and the coherence of the association. The report that followed was an extended attempt to review all the evidence then available and synthesize it into an overall judgment: smoking causes lung cancer, bronchitis, emphysema, and is a "health hazard of sufficient importance in the United States to warrant appropriate remedial action" (Bayne-Jones et al., 1964, p. 33).

Although the Surgeon General's 1964 report did not explicitly categorize the *level* of evidential support for any of its conclusions, the 2004 Surgeon General's report does, as do recent reports from IARC (2006b) and IOM (2006a). A variety of categorizations exist. For example, the 2004 Surgeon General's report on smoking (DHHS/CDC, 2004), as well as the 2006 IOM Committee on Asbestos (IOM, 2006a, p. 20), employed a four-level categorization of the strength of evidence of causation (the latter was based on the 2004 Surgeon General's Report on Smoking [DHHS/CDC, 2004]):

1. Sufficient to infer a causal relationship
2. Suggestive but not sufficient to infer a causal relationship
3. Inadequate to infer the presence or absence of a causal relationship
4. Suggestive of no causal relationship

IARC forms expert committees and instructs them to first categorize the level of evidence within three subcategories—human, animal, and mechanistic—and then to synthesize the subcategories of evidence into an overall evaluation on a five-category scale ranging from carcinogenic to probably not carcinogenic. Figure 8-3 depicts the IARC evaluation scheme.

Several groups, including some IOM committees on Agent Orange (IOM, 1994, 1996, 1999, 2001, 2003b, 2005b) and other veterans' health issues, have based their approaches on the IARC (2006b) system for classifying the evidence in the *subcategory* pertaining only to human evidence. Within this subcategory, IARC systematically reviews available epidemiologic studies, considering study quality, relevance, and strength of findings. It then classifies the overall epidemiologic evidence as *sufficient evidence* if there is a finding that "a positive relationship has been observed between the exposure and cancer in studies in which chance, bias, and confounding could be ruled out with reasonable confidence" (IARC, 2006b, p. 19). For *limited evidence*, a "positive association has been observed between exposure to the agent and cancer for which a causal interpretation is considered by the working group to be credible, but chance, bias, or confounding could not be ruled out with reasonable confidence" (IARC, 2006b, pp. 19-20). With *inadequate evidence*, either there are no epidemiologic data or study quality, or power, or consistency across studies precludes a conclusion regarding causal association.

The IOM committees assessing the impact of Agent Orange have published biennial reports since 1994. Table 8-1 contains the four-level categorization for the strength of epidemiologic evidence used in the 1994 report (IOM, 1994), which is quite similar to IARC's subcategory scheme for human evidence (IARC, 2006b). Subsequent reports have used a similar classification scheme.

IOM committees (IOM, 2000, 2003a, 2004, 2005a, 2006b, 2007) examining Gulf War and health have added a causal category to those used by the IOM Agent Orange committees (IOM, 1994, 1996, 1999, 2001, 2003b, 2005b). Table 8-2 shows the classification scheme from Volume 1 of the Gulf War and Health series (IOM, 2000). The additional category makes causation explicit and includes evidence beyond that found in just epidemiologic studies.

The IOM Agent Orange categorization (IOM, 1994, 1996, 1999, 2001, 2003b, 2005b) relies less on mechanistic and animal evidence and more
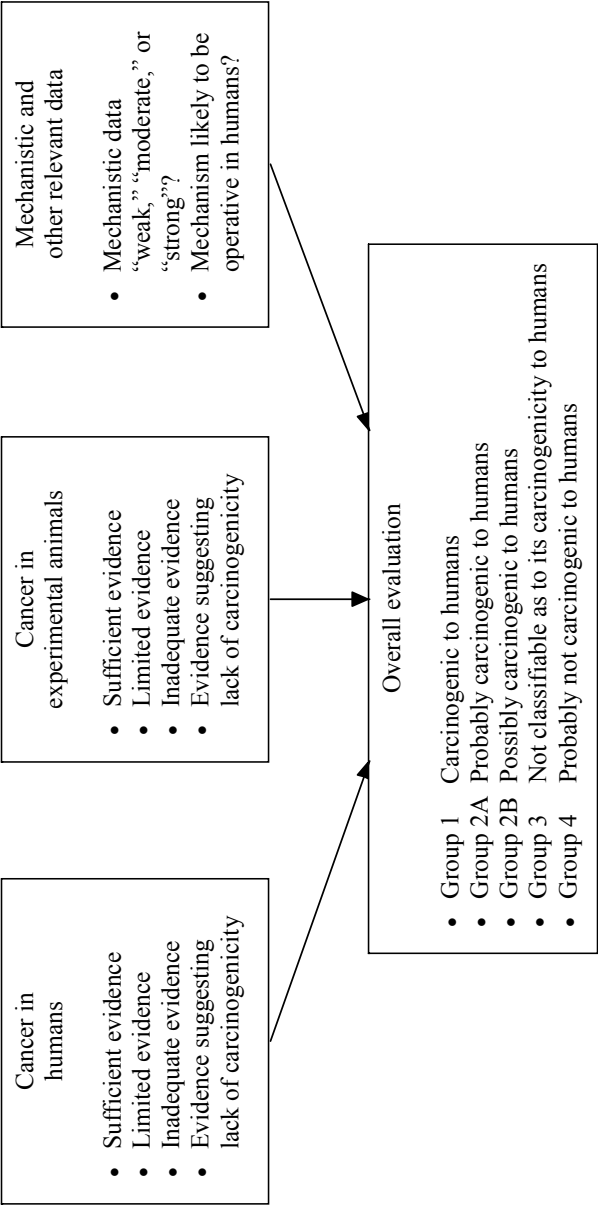
FIGURE 8-3 IARC evaluation scheme.
SOURCE: Adapted from Cogliano, 2006.

TABLE 8-1 IOM Categorization from the Executive Summary of *Veterans and Agent Orange: Health Effects of Herbicides Used in Vietnam*

Sufficient Evidence of an Association	Evidence is sufficient to conclude that there is a positive association. That is, a positive association has been observed between herbicides and the outcome in studies in which chance, bias, and confounding could be ruled out with reasonable confidence. For example, if several small studies that are free from bias and confounding show an association that is consistent in magnitude and direction, there may be sufficient evidence of an association.
Limited/Suggestive Evidence of an Association	Evidence is suggestive of an association between herbicides and the outcome but is limited because chance, bias, and confounding could not be ruled out with confidence. For example, at least one high-quality study shows a positive association, but the results of other studies are inconsistent.
Inadequate/Insufficient Evidence to Determine Whether an Association Exists	The available studies are of insufficient quality, consistency, or statistical power to permit a conclusion regarding the presence or absence of an association. For example, studies fail to control for confounding, have inadequate exposure assessment, or fail to address latency.
Limited/Suggestive Evidence of No Association	Several adequate studies, covering the full range of levels of exposure that human beings are known to encounter, are mutually consistent in not showing a positive association between exposure to herbicides and the outcome at any level of exposure. A conclusion of "no association" is inevitably limited to the conditions, level of exposure, and length of observation covered by the available studies. *In addition, the possibility of a very small elevation in risk at the levels of exposure studied can never be excluded.*

SOURCE: IOM, 1994.

on epidemiologic data than do the IARC, NTP, or EPA models. As there may be many cases relevant to presumptive service connection in which the epidemiologic evidence on veterans specifically is extremely thin or even non-existent, giving animal and mechanistic evidence a more prominent and systematic role in the overall evaluation scheme is warranted. The causal category added by the Gulf War committees is a step in this direction. Previous IOM Agent Orange committees also reviewed and reported on toxicological and mechanistic information, but according to their interpretation of their charge, did not figure this information into the conclusions about the strength of evidence for association and causation.

TABLE 8-2 IOM Categorization from the Executive Summary of *Gulf War and Health, Volume 1: Depleted Uranium, Pyridostigmine Bromide, Sarin, Vaccines*

Sufficient Evidence of a Causal Relationship	Evidence is sufficient to conclude that a causal relationship exists between the exposure to a specific agent and a health outcome in humans. The evidence fulfills the criteria for sufficient evidence of an association (below) and satisfies several of the criteria used to assess causality: strength of association, dose-response relationship, consistency of association, temporal relationship, specificity of association, and biological plausibility.
Sufficient Evidence of an Association	Evidence is sufficient to conclude that there is a positive association. That is, a positive association has been observed between an exposure to a specific agent and a health outcome in human studies in which chance, bias, and confounding could be ruled out with reasonable confidence.
Limited/Suggestive Evidence of an Association	Evidence is suggestive of an association between exposure to a specific agent and a health outcome in humans, but is limited because chance, bias, and confounding could not be ruled out with confidence.
Inadequate/Insufficient Evidence to Determine Whether an Association Does or Does Not Exist	The available studies are of insufficient quality, consistency, or statistical power to permit a conclusion regarding the presence or absence of an association between an exposure to a specific agent and a health outcome in humans.
Limited/Suggestive Evidence of No Association	There are several adequate studies covering the full range of levels of exposure that humans are known to encounter, that are mutually consistent in not showing a positive association between exposure to a specific agent and a health outcome at any level of exposure. A conclusion of no association is inevitably limited to the conditions, levels of exposure, and length of observation covered by the available studies. In addition, the possibility of a very small elevation in risk at the levels of exposure studied can never be excluded.

SOURCE: IOM, 2000.

A PROPOSED QUALITATIVE FRAMEWORK FOR EVALUATING CAUSAL CLAIMS

Incorporating the Full Range of Evidence

The new process recommended by this Committee involves a categorization of the strength of evidence in support of a causal claim that incorporates the full weight of all evidence, including expert opinion, findings from

epidemiologic and animal studies, and mechanistic knowledge. Reliance on the broad range of pertinent scientific data is also in keeping with the original congressional language that specifies consideration by NAS of the biological plausibility of any association. Before describing our proposed classification scheme, we briefly consider the consequences of extending the range of evidence considered. One issue is whether a seemingly more rigorous bar of evidence would reduce the likelihood of reaching a classification level at which compensation is made.

We approach this issue through case studies. Incorporating information derived from mechanistic studies and animal toxicology can lead to an upgrading or downgrading of a classification based on the weight of evidence classification. Some examples are described below.

Particulate Air Matter

Deployed personnel in the first Gulf conflict were exposed to airborne particles from the oil fires in Kuwait, exhaust from military vehicles and other combustion sources, and dust stirred by troop movements. The health effects of airborne particles are of general concern and the scientific evidence is subject to periodic review in the setting of the National Ambient Air Quality Standard (NAAQS) by the U.S. Environmental Protection Agency (SOURCE: http://epa.gov/pm/naaqsrev2006.html). The standard is evidence-based and the process of setting a new NAAQS involves a comprehensive review of all relevant scientific evidence, including epidemiologic and toxicological data as well as information on exposure patterns. For the last two reviews of the NAAQS, 1996 and 2006, the epidemiological evidence has been extensive, showing significant associations of airborne particles with increased risk for premature mortality and morbidity. The effects are relatively small, however, and biological plausibility has been a major consideration in evaluating the evidence and determining if the associations can be judged as causal. Information relevant to the plausibility of the associations comes from studies of the chemical and physical properties of particles and from toxicological studies that have addressed responses to particles in in vitro and in vivo models. Using a "weight of evidence" approach, the Environmental Protection Agency has judged the associations of airborne particles with adverse effects to be causal (SOURCE: http://epa.gov/pm/naaqsrev2006.html).

2,3,7,8-Tetrachlorodibenzo-para-dioxin (TCDD)

TCDD, a potent dioxin, was listed in 1997 as a Group 1 carcinogen by IARC based on limited evidence in humans, sufficient evidence in experimental animals, and abundant mechanistic information including data

demonstrating that TCDD acts through the aryl hydrocarbon receptor (AhR), which is present in both humans and animals (IARC, 1997).

Formaldehyde

IARC has recently concluded that formaldehyde should be added to that group of agents that are carcinogenic to humans (Group 1 carcinogen) (IARC, 2004, p. 1). This upgrade was based upon new epidemiologic evidence of an association with nasopharyngeal cancer. "[T]here is now sufficient evidence that formaldehyde causes nasopharyngeal cancer in humans, a rare cancer in developed countries" (IARC, 2004, p. 1). Previous laboratory, animal, and mechanistic evidence supported this association. The IARC found "strong but not sufficient evidence for a causal association between leukemia and occupational exposure to formaldehyde," falling "slightly short of being fully persuasive because" of limitations in the cohort and conflict with nonpositive findings in another cohort (IARC, 2006a, p. 5). IARC noted findings of lymphomas and leukemias in one study in male rats, and several possible mechanisms for the induction of human leukemia, such as clastogenic damage to circulatory stem cells. However IARC also noted the lack of good rodent models that simulate the occurrence of acute myeloid leukemia in humans, and did not identify a mechanism for its induction in humans (IARC, 2006a).

Saccharin

Saccharin was originally classified as "reasonably anticipated to be a human carcinogen" by NTP (NIH, 2000, p. 1) and "possibly carcinogenic to humans" by IARC (IARC, 1987, p. 334) based on data clearly showing bladder cancer in rats. Although there was some limited epidemiologic evidence associating bladder cancer with saccharin sweeteners, the epidemiologic evidence was classified by IARC as inadequate. Subsequent mechanistic studies attributed the animal cancer findings as being due to a mechanism that would only occur at high doses in rats. "Saccharin produces urothelial bladder tumours in rats by a non-DNA-reactive mechanism that involves the formation of a urinary calcium phosphate-containing precipitate, cytotoxicity and enhanced cell proliferation. This mechanism is not relevant to humans because of critical interspecies differences in urine composition" (IARC, 1999, p. 50). Saccharin has since been delisted as a reasonably anticipated carcinogen by NTP and deemed "not classifiable as to its carcinogenicity to humans" by IARC (IARC, 1999, p. 50; NIH, 2000).

These case studies illustrate the potential contributions of mechanistic information as evidence relevant to causation is evaluated. A certain under-

standing of mechanism of action may have substantial impact in considering the overall weight of evidence.

Committee Recommended Categories for the Level of Evidence for Causation

In light of the categorizations used by other health organizations and agencies as well as considering the particular challenges of the presumptive disability decision-making process, we propose a four-level categorization of the strength of the *overall evidence* for or against a *causal relationship* from exposure to disease:

1. *Sufficient*: The evidence is sufficient to conclude that a causal relationship exists.
2. *Equipoise and Above*: The evidence is sufficient to conclude that a causal relationship is at least as likely as not, but not sufficient to conclude that a causal relationship exists.
3. *Below Equipoise*: The evidence is not sufficient to conclude that a causal relationship is at least as likely as not, or is not sufficient to make a scientifically informed judgment.
4. *Against*: The evidence suggests the lack of a causal relationship.

We use the term "equipoise" to refer to the point at which the evidence is in balance between favoring and not favoring causation. The term "equipoise" is widely used in the biomedical literature, is a concept familiar to those concerned with evidence-based decision making, and is used in VA processes for rating purposes as well as being a familiar term in the veterans' community.

Below we elaborate on the four-level categorization that the Committee recommends.

Sufficient

If the overall evidence for a causal relationship is categorized as Sufficient, then it should be scientifically compelling. It might include

- replicated and consistent evidence of a causal association: that is, evidence of an association from several high-quality epidemiologic studies that cannot be explained by plausible noncausal alternatives (e.g., chance, bias, or confounding), or
- evidence of causation from animal studies and mechanistic knowledge, or
- compelling evidence from animal studies and strong mechanistic

evidence from studies in exposed humans, consistent with (i.e., not contradicted by) the epidemiologic evidence.

Using the Bayesian framework to illustrate the evidential support and the resulting state of communal scientific opinion needed for reaching the Sufficient category (and the lower categories that follow), consider again the causal diagram in Figure 8-2. In this model, used to help clarify matters conceptually, the observed association between exposure and health is the result of (1) measured confounding, parameterized by α; (2) the causal relation, parameterized by β; and (3) other, unmeasured sources such as bias or unmeasured confounding, parameterized by γ. The belief of interest, after all the evidence has been weighed, is in the size of the causal parameter β. Thus, for decision making, what matters is how strongly the evidence supports the proposition that β is above 0. As it is extremely unlikely that the types of exposures considered for presumptions reduce the risk of developing disease, we exclude values of β below 0. If we consider the evidence as supporting degrees of belief about the size of β, and we have a posterior distribution over the possible size of β, then a posterior like Figure 8-4 illustrates a belief state that might result when the evidence for causation is considered Sufficient.

As the "mass" over a positive effect (the area under the curve to the right of the zero) vastly "outweighs" the small mass over no effect (zero), the evidence is considered sufficient to conclude that the association is causal. Put another way, even though the scientific community might be uncertain as to the size of β, after weighing all the evidence, it is highly confident that the probability that β is greater than zero is substantial; that is, that exposure causes disease.
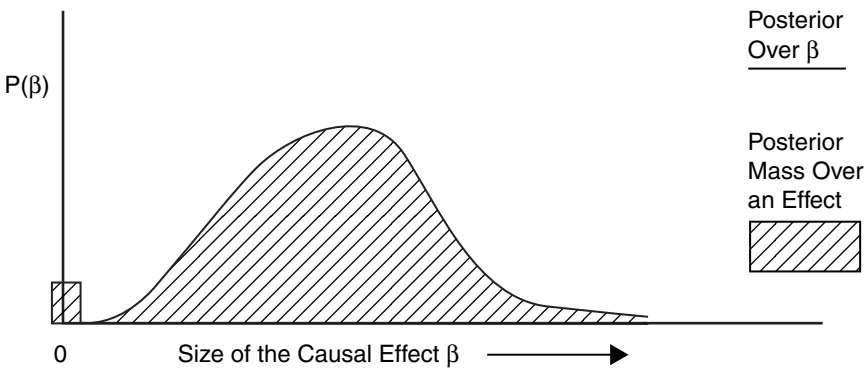

FIGURE 8-4 Example posterior for *Sufficient*.

Equipoise and Above

To be categorized as Equipoise and Above, the scientific community should categorize the overall evidence as making it more confident in the existence of a causal relationship than in the non-existence of a causal relationship, but not sufficient to conclude causation.

For example, if there are several high-quality epidemiologic studies, the preponderance of which show evidence of an association that cannot readily be explained by plausible noncausal alternatives (e.g., chance, bias, or confounding), and the causal relationship is consistent with the animal evidence and biological knowledge, then the overall evidence might be categorized as Equipoise and Above. Alternatively, if there is strong evidence from animal studies or mechanistic evidence, not contradicted by human or other evidence, then the overall evidence might be categorized as Equipoise and Above. Equipoise is a common term employed by VA and the courts in deciding disability claims (see Appendix D).

Again, using the Bayesian model to illustrate the idea of Equipoise and Above, Figure 8-5 shows a posterior probability distribution that is an example of belief compatible with the category Equipoise and Above.

In this figure, unlike the one for evidence classified as Sufficient, there is considerable mass over zero, which means that the scientific community has considerable uncertainty as to whether exposure causes disease at all; that is, whether β is greater than zero. At *least* half of the mass is to the right of the zero, however, so the community judges causation to be at least as likely as not, after they have seen and combined all the evidence available.
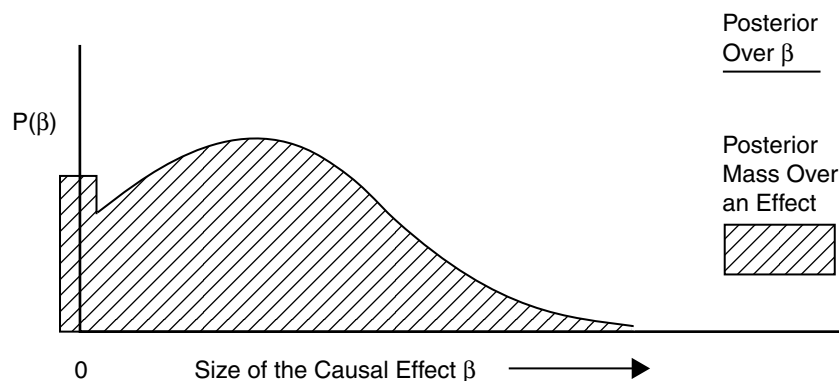

FIGURE 8-5 Example posterior for *Equipoise and Above*.

Below Equipoise

To be categorized as Below Equipoise, the overall evidence for a causal relationship should either be judged not to make causation at least as likely as not, or not sufficient to make a scientifically informed judgment.

This might occur

1. when the human evidence is consistent in showing an association, but the evidence is limited by the inability to rule out chance, bias, or confounding with confidence, and animal or mechanistic evidence is weak, or

2. when animal evidence suggests a causal relationship, but human and mechanistic evidence is weak or inconsistent, or

3. when mechanistic evidence is suggestive but animal and human evidence is weak or inconsistent, or

4. when the evidence base is very thin.

Figure 8-6 shows a posterior probability distribution that is an example of belief compatible with the category Below Equipoise.

Against

To be categorized as Against, the overall evidence should favor belief that there is no causal relationship from exposure to disease. For example, if there is human evidence from multiple studies covering the full range of exposures encountered by humans that are consistent in showing no causal association, or there is animal or mechanistic evidence supporting the lack of a causal relationship, and combining all of the evidence results
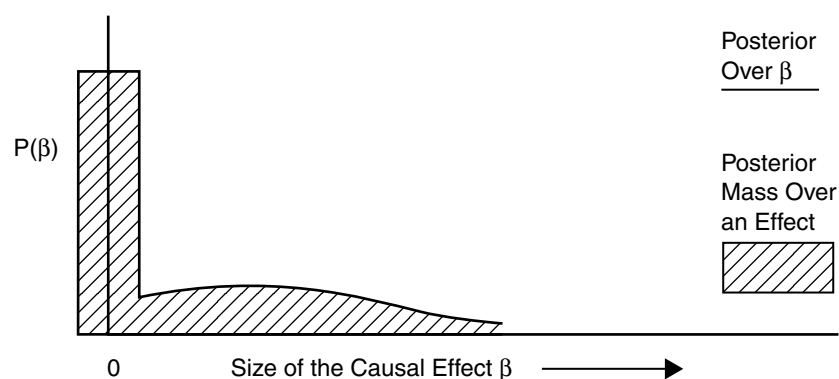

FIGURE 8-6 Example posterior for *Below Equipoise*.

in a posterior resembling Figure 8-7, then the scientific community should categorize the evidence as *Against* causation.

Comparison of the Committee's Proposed and Previous Frameworks

The Committee's proposed framework departs from that used by previous IOM committees in assessing Agent Orange (Table 8-1) in at least three respects. First, as noted, previous IOM committees evaluating Agent Orange (IOM, 1994, 1996, 1999, 2001, 2003b, 2005b) relied primarily on epidemiologic studies to classify strength of evidence for association and did not systematically incorporate evidence from animal toxicology and mechanistic studies.

Second, the Agent Orange categorization differentiated between levels of evidence for *association* instead of *causation* (IOM, 1994, 1996, 1999, 2001, 2003b, 2005b). As we have proposed in the previous chapter, the claims at issue related to compensation of veterans are causal claims, not associational claims. Association, especially association adjusted for potential confounders, is *evidence* for the causal claim, but it is *not identical to* the causal claim. In fact, making a presumptive decision to compensate on the basis of "limited/suggestive" evidence of an association presents the possibility that there is no causal link at all, and all of those receiving compensation because of such a presumption could be false positives. A causal category was added by previous IOM committees on Gulf War and Health, but the categories below "Sufficient evidence of a causal relationship" retain the language of association (IOM 2000, 2003, 2004, 2005a, 2006b, 2007). This Committee recommends, therefore, a categorization
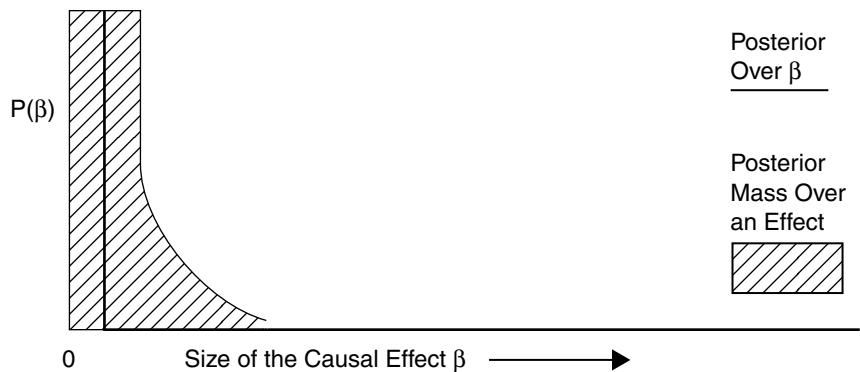

FIGURE 8-7 Example posterior for *Against*.

that makes explicit the evidential role of association, but that keeps the clear and explicit overall goal of assessing causation.

Third, we created the category of Equipoise and Above to capture the spirit of presumption: the tie goes to the veteran, and to stay true and scientifically consistent to the evidential standard suggested by Congress in the Agent Orange Act of 1991 (Public Law Number 102-4, 102d Cong., 1st Sess.), which gave the VA Secretary authority to prescribe regulations providing for a presumption "[w]henever the Secretary determines, on the basis of sound medical and scientific evidence, that a positive association exists between" herbicide exposure and a disease. Section (4)(b)(3) states the following:

> An association between the occurrence of a disease in humans and exposure to an herbicide agent shall be considered to be positive for the purposes of this section if the credible evidence for the association is equal to or outweighs the credible evidence against the association.

In our categorization, Equipoise and Above represents a state in which there *is* credible evidence, and the credible evidence *for* causation is equal to or greater than the credible evidence *against* causation. We also intend this categorization to be flexible over time. We expect that, as the evidence base grows, evaluations about the state of evidence for a causal claim may be upgraded *or* downgraded over time. The descriptive categorization language used by previous IOM Agent Orange committees for "limited/suggestive" evidence of an association implies that a single high-quality epidemiologic study can in some circumstances be sufficient for the "limited/suggestive" category:

> Evidence is suggestive of an association between herbicides and the outcome but is limited because chance, bias, and confounding could not be ruled out with confidence. For example, at least one high-quality study shows a positive association, but the results of other studies are inconsistent. (IOM, 1994, p. 97; IOM, 1996, 1999, 2001, 2003b, 2005b)

If a scientific committee's conclusion were based on a single study and later studies were to show more definitive evidence of an association, a subsequent committee could upgrade the "limited/suggestive" classification of this association to Sufficient. On the other hand, if definitive studies were reported that supported an overall weight of the evidence of below "limited/ suggestive," a subsequent committee could downgrade the classification. Under the current approach, however, it is unclear if any reclassifications of evidence may lead to a change in a presumptive decision that VA has established based upon the classification of "limited/suggestive" evidence of an association.

SUMMARY

Combining human evidence with meta-analysis is very useful when several reasonably exchangeable studies are to be combined. However, the technique cannot be used to combine the full range of evidence relevant to classifying the level of evidence for causation. A very general technique for combining diverse evidence into a single, quantitative description of belief about the causal claim at stake is the Bayesian approach, but its usefulness in a presumptive disability decision-making context may be limited. As a result, other organizations such as IARC have resorted to a qualitative categorization of the strength of evidence for causation. These agencies base the overall categorization on separate judgments about the strength of evidence from epidemiologic studies, animal studies, or other mechanistic, toxicological, or biological sources.

For the presumptive disability decision-making process, this Committee recommends categorizing the level of overall evidence for a causal relationship between exposure and health outcome in one of the following categories:

1. *Sufficient*: The evidence is sufficient to conclude that a causal relationship exists.

2. *Equipoise and Above*: The evidence is sufficient to conclude that a causal relationship is at least as likely as not, but not sufficient to conclude that a causal relationship exists.

3. *Below Equipoise*: The evidence is not sufficient to conclude that a causal relationship is at least as likely as not, or is not sufficient to make a scientifically informed judgment.

4. *Against*: The evidence suggests the lack of a causal relationship.

REFERENCES

Bayne-Jones, S., W. J. Burdette, W. G. Cochran, E. Farber, L. F. Fieser, J. Furth, J. B. Hickam, C. LeMaistre, L. M. Schuman, and M. H. Seevers. 1964. *Smoking and health. Report of the Advisory Committee to the Surgeon General of the Public Health Service*. Public Health Services Publication No. 1103. Washington, DC: Public Health Service. http://www.cdc.gov/tobacco/sgr/sgr_1964/sgr64.htm (accessed February 13, 2007).

Berlin, J. A., and E. M. Antman. 1994. Advantages and limitations of metaanalytic regressions of clinical trials data. *Online Journal of Current Clinical Trials*. Doc. No. 134.

Berlin, J., and T. C. Chalmers. 1988. Commentary on meta-analysis in clinical trials. *Hepatology* 8:690-691.

Cogliano, V. J. 2006. *IARC monographs on the evaluation of carcinogenic risks to humans*. Slideshow. http://www.oehha.ca.gov/prop65/public_meetings/pdf/IARCVJCogliano121206.pdf. (accessed May 14, 2007).

DHHS/CDC (Department of Health and Human Services/Centers for Disease Control and Prevention). 2004. The health consequences of smoking: A report of the Surgeon General. http://www.cdc.gov/tobacco/data_statistics/sgr/sgr_2004/index.htm (accessed July 27, 2007).

Dickersin, K., and J. A. Berlin. 1992. Meta-analysis: State-of-the-science. *Epidemiology Review* 14:154-176.

EPA (Environmental Protection Agency). 2005. *Guidelines for carcinogen risk assessment*. EPA/630/P-03/001F. Washington, DC: EPA Risk Assessment Forum.

Greenland, S. 1994a. Can meta-analysis be salvaged? *American Journal of Epidemiology* 140:783-787.

Greenland, S. 1994b. Invited commentary: A critical look at some popular meta-analytic methods. *American Journal of Epidemiology* 140:290-296.

Greenland, S., and K. O'Rourke. 2001. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. *Biostatistics* 2:463-471.

Howson, C., and P. Urbach. 1989. *Scientific reasoning: The Bayesian approach*. Peru, IL: Open Court Publishing Company.

IARC (International Agency for Research on Cancer). 1987. *IARC monographs on the evaluation of carcinogenic risks to humans overall evaluations of carcinogenicity: An updating of IARC Monographs*. Vol. 1 to 42. Supplement 7. http://monographs.iarc.fr/ENG/Monographs/suppl7/suppl7.pdf (accessed May 4, 2007).

IARC. 1997. *Polychlorinated dibenzo-para-dioxins and polychlorinated dibenzofurans. IARC monographs on the evaluations of carcinogenic risks to humans.* Vol. 69. http://monographs.iarc.fr/ENG/Monographs/vol69/volume69.pdf (accessed May 4, 2007).

IARC. 1999. IARC *Monographs on the evaluation of carcinogenic risks to humans. Some chemicals that cause tumours of the kidney or urinary bladder in rodents and some other substances. Summary of data reported and evaluation.* Vol. 73. http://monographs.iarc.fr/ENG/Monographs/vol73/volume73.pdf (accessed May 4, 2007).

IARC. 2004. *IARC classifies formaldehyde as carcinogenic to humans.* Press Release No. 153. http://www.iarc.fr/ENG/Press_Releases/archives/pr153a.html (accessed February 13, 2007).

IARC. 2006a. *Formaldehyde, 2-butoxyethanol and 1-tert-butoxypropan-2-ol.* IARC monograph. Vol. 88. http://monographs.iarc.fr/ENG/Monographs/vol88/volume88.pdf (accessed February 13, 2007).

IARC. 2006b. *IARC monographs on the evaluation of carcinogenic risks to humans: Preamble.* Lyon, France: World Health Organization/International Agency for Research on Cancer. http://monographs.iarc.fr/ENG/Preamble/CurrentPreamble.pdf (accessed February 13, 2007).

IOM (Institute of Medicine). 1994. *Veterans and Agent Orange: Health effects of herbicides used in Vietnam.* Washington, DC: National Academy Press.

IOM. 1996. *Veterans and Agent Orange: Update 1996.* Washington, DC: National Academy Press.

IOM. 1999. *Veterans and Agent Orange: Update 1998.* Washington, DC: National Academy Press.

IOM. 2000. *Gulf War and health, volume 1: Depleted uranium, pyridostigmine bromide, sarin, vaccines.* Washington, DC: National Academy Press.

IOM. 2001. *Veterans and Agent Orange: Update 2000.* Washington, DC: National Academy Press.

IOM. 2003a. *Gulf War and health, volume 2: Insecticides and solvents.* Washington, DC: The National Academies Press.

IOM. 2003b. *Veterans and Agent Orange: Update 2002.* Washington, DC: The National Academies Press.

IOM. 2004. *Gulf War and health: Updated literature review of sarin.* Washington, DC: The National Academies Press.

IOM. 2005a. *Gulf War and health, volume 3: Fuels, combustion products, and propellants.* Washington, DC: The National Academies Press.

IOM. 2005b. *Veterans and Agent Orange: Update 2004*. Washington, DC: The National Academies Press.

IOM. 2006a. *Asbestos: Selected cancers*. Washington, DC: The National Academies Press.

IOM. 2006b. *Gulf War and health, volume 4: Health effects of serving in the Gulf War*. Washington, DC: The National Academies Press.

IOM. 2007. *Gulf War and health, volume 5: Infectious diseases*. Washington, DC: The National Academies Press.

IOM/NRC (National Research Council). 2005. *Dietary supplements: A framework for evaluating safety*. Washington, DC: The National Academies Press.

Kaizar, E. E. 2005. Meta-analyses are observational studies: How lack of randomization impacts analysis. *American Journal of Gastroenterology* 100:1233-1236.

NAS (National Academy of Sciences). 1988. *Health risks of radon and other internally deposited alpha-emitters (Beir IV)*. Washington, DC: National Academy Press.

NIH (National Institutes of Health). 2000. *Fact sheet: The report on carcinogens.* 9th ed. http://www.nih.gov/news/pr/may2000/niehs-15.htm (accessed June 3, 2007).

NTP (National Toxicology Program). 2005. *Report on carcinogens. 11th report on Carcinogens.* Research Triangle Park, NC: NTP.

NTP CERHR (NTP Center for the Evaluation of Risks to Human Reproduction). 2003. *NTP-CERHR Monograph on the Potential Human Reproductive and Developmental Effects of Di-n-Butyl Phthalate (DBP)*. Research Triangle Park, NC: NTP CERHR, U.S. Department of Health and Human Services.

NTP CERHR. 2005. *Guidelines for CERHR Expert Panel Members*. Research Triangle Park, NC: NTP CERHR, U.S. Department of Health and Human Services.

Shelby, M.D. 2005. National Toxicology Program Center for the Evaluation of Risks to Human Reproduction: Guidelines for CERHR Expert Panel Members. *Birth Defects Research (Part B)* 74:9-16.

Stram, D. O. 1996. Meta-analysis of published data using a linear mixed-effects model. *Biometrics* 52:536-544.

Stroup, D. F., J. A. Berlin, S. C. Morton, I. I. Olkin, G. D. Williamson, D. Rennie, D. Moher, B. J. Becker, T. A. Sipe, and S. B. Thacker. 2000. Meta-analysis of observational studies in epidemiology: A proposal for reporting. Meta-analysis of Observational Studies in Epidemiology (MOOSE) group. *Journal of the American Medical Association* 283: 2008-2012.